图书情报工作
第 64 卷 第 11 期　2020 年 6 月

依存句法特征的科研命名实体识别算法*

■ 赵华茗[1]　钱力[1,2]　余丽[1]

[1]中国科学院文献情报中心　北京 100190　[2]中国科学院大学经济与管理学院图书情报与档案管理系　北京 100190

摘　要：[目的/意义]探索科研命名实体及其关系的识别与抽取，提升其在长句等复杂情况下的识别效果，为进一步的应用提供参考与借鉴。[方法/过程]以依存句法特征分析为基础，提出一种科研命名实体关系抽取方法，过程包括：①使用 Standford Tagger 工具对目标文本进行词性标注；②基于标注结果，围绕核心谓词和 SAO 结构，将目标文本分割为结构规范的语义片段；③通过依存句法分析，找出与核心谓词语义相关的主语和宾语，构成(实体，关系，实体)三元组。[结果/结论]与 Ollie、Reverb 等主流算法进行的对比测试表明，该方法可以有效提升科研命名实体识别的准确性。

关键词：依存句法分析　科研命名实体　实体识别　关系抽取
分类号：G250
DOI：10.13266/j.issn.0252-3116.2020.11.012

1　引言

大数据时代，如何从海量的数据中获取有用信息成为自然语言处理和数据挖掘中的难点和热点问题。作为自然语言处理的基础之一，实体及实体关系抽取为词法、句法、语义等分析处理提供了重要技术支撑，广泛用于信息抽取、信息检索、信息推荐、分类聚类、自动文摘、自动问答、知识发现、情感分析、知识库构建等众多自然语言处理任务中。

针对实体关系复杂的情况，徐芬等[1]提出基于特征向量的实体及实体关系抽取方法，融合了词、词性标注、实体属性、实体间关系等特征信息，他们的研究表明多个层次的语言学特征能够有效提升实体关系抽取的效果。N. Kambhatla[2]融合了实体单词、实体类型、实体引用方式、重叠、依存树和解析树等特征信息，基于最大熵模型实现实体及关系抽取。郭喜跃等[3]在句法分析的基础上，提出句法与语义特征融合的实体关系抽取方法，该方法主要融合了句法依存关系、实体与核心谓词的距离、语义角色标注等信息，可以有效识别实体间的多种关系。甘丽新等[4]在此基础上融入了依存句法组合特征及动词依赖，使识别的关系类型种类有了很大的提高。H. Li 等[5]提出了一种基于位置语义特征的实体关系抽取方法，利用位置特征的可计算性和可操作性，以及语义特征的可理解性和可实现性，整合了词语位置的信息增益与基于 HowNet 的语义计算结果。实验结果表明，结合位置和语义特征的关系抽取方法优于单独使用位置或语义特征的方法。奚斌等[6]也是通过在各种词法、语法、语义的基本特征内部及特征之间进行有效的组合，形成多种组合特征来提高实体关系的抽取性能和效果。

以上研究表明，融合句法依存和词性标注信息能够有效地提高实体关系抽取的性能。常见的抽取工具中，TextRunner[7]，Reverb[8]和 R2A2[9]利用句法分析算法实现信息抽取，而 WOE[10]，KrakeN[11]，Ollie[12]等则进一步融合了句法依存分析算法，抽取效果更好。这些工具设计思路皆是通过对文本中表达关系短语的模式进行仔细的语言分析后，形成模式集，然后再结合正则表达式和模式匹配算法，实现高精度的实体及关系抽取[13]。近年来，基于深度学习的实体关系抽取技术的研究[14]也取得了相应的成果。唐敏等[15]通过增加实体注意力机制的深度学习实体关系抽取模型来辨别语义关系；Y. Lin 等[16]提出了一种在纯文本中进行关

* 本文系中国科学院文献情报能力建设专项项目“文献情报‘数据湖’及开放式大数据框架建设”(项目编号：院 1852)与国家科技图书文献中心专项任务“多源数据增值与知识计算方法研究”(项目编号：K180201001)研究成果之一。

作者简介：赵华茗(ORCID：0000-0002-8829-9208)，副研究馆员，E-mail：zhaohm@mail.las.ac.cn；钱力(ORCID：0000-0002-0931-2882)，副研究馆员；余丽(ORCID：0000-0002-4374-8743)，馆员。

收稿日期：2019-09-24　修回日期：2020-02-02　本文起止页码：108-115　本文责任编辑：杜杏叶

系抽取的方法,在引入多语言的神经关系抽取框架的基础上也加入了注意机制,有效地控制了噪声句子的影响。

本文提出一种科研命名实体关系抽取方法,采用词性标注和句法依存相融合的方式直接分析句子,获得最终的实体关系三元组。相较于上文中提到的通过正则表达式及模式匹配提取实体关系三元组的方式,该方法在如下两个方面进行了改进:①长句处理方面,围绕核心谓词和 SAO 结构,将长句分割为结构规范的语义片段,以利于下一步的实体对的准确抽取与识别;②科研命名实体关系识别方面,通过对核心谓词与其辅助词的依存关系分析建模,对科研命名实体识别模型进行了优化,有效提升识别准确性,如“To efficiently handle high-dimensional data, we develop two deterministic algorithms that approximate the covariance matrices.”句子中,Ollie 等工具只能识别出围绕核心动词“develop”的实体关系(we; develop; two deterministic algorithms)。从句子本身含义来看,科研命名实体候选三元组的结果更希望是“developed A to handle B”中科研命名实体 A 和 B 的关系,即(two deterministic algorithms; be developed to; high-dimensional data)。本文通过增加动词“develop”和辅助词“to”的句法依存分析算法,实现了对科研论文中重要科研命名实体及其关系的抽取。

2 实体关系抽取算法

2.1 算法设计思路

词性标注和依存句法特征的实体关系抽取算法整体设计思路如图 1 所示:

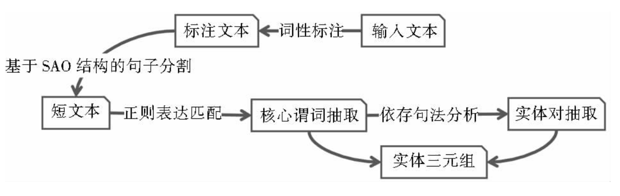


图 1　实体及关系抽取算法架构设计

算法原理如下:

(1) 以 SAO 结构作为基本句子结构,进行长句切割化简处理。首先利用 Standford NLP 工具,对输入文本进行词性标注,然后依据 SAO 基本结构单元,通过句法分析,找到核心谓词,并围绕核心谓词和 SAO 结构单元,将长句分割为更为细粒度的语义结构单元。

(2) 通过依存句法特征分析,建模,找出与核心谓词语义相关的主语和宾语。

(3) 整合主语、核心谓词和宾语构成[实体,关系,实体]三元组。

2.2　基于 SAO 结构的长句切割

SAO(Subject-Action-Object)结构理论,源自于创造性问题解决理论(Theory of Inventive Problem Solving, TIPS),用来表示解决问题方法的基本函数单元[17]。从句子语法结构看,SAO 结构可以对应句子中的 SVO(Subject-Verb-Object)结构;从语义网 RDF 数据模型看,SAO 结构可以对应三元组(triple)中的 SPO(Subject-Predication-Object)结构。SAO 结构的引入可以有效揭示组件信息和组件间的语义关系[18],进而形成一个完整的语义理解。近年来,SAO 结构广泛应用于技术路线分析[19]、技术演化[20]等语义分析领域。相较基于句子的分析,SAO 结构提供了一种更为细粒度的语义结构,有助于更为深入、更为准确地挖掘和理解文本中蕴含的关联信息。

针对长句的实体抽取问题,A. Gabor[21]依据规范结构(Canonically Structured),先将长句分解为一组短句,然后以自然逻辑推理方式从短句中确定候选三元组。L. Corro[22]围绕 7 个基本句型将长句分解为一组短句,再通过依存关系分析从短句中确定候选三元组,提升抽取效果。研究发现,无论规范结构或是基本句型都包含核心谓词部分,因此,本文以核心谓词为单元,以 SAO 结构为基本结构单元和校验模型将长句分割为更细粒度的语义结构,是合理的,有利于下一步的规范语义结构中实体的精准抽取。本文使用长句分割方式,而没有使用长句分解短句方式,主要有如下两个方面的考虑:①实体识别过程中的句法依存分析,可以直接利用长句的句法依存分析结果,减少中间过程,减小错误率;②满足 SAO 结构的基础上,尽量保留原句的信息,减少信息丢失。

基于 SAO 结构的长句切割的实现过程包括如下四步:

(1)利用 Standford Tagger 工具对长句进行词性标注,标注结果中,可以看出名词以“NP”为起始标识,而动词以“VP” 为起始标识,规律明显;

(2)对词性标注结果进行预处理,主要是对不定式等非谓语动词进行特征标识,以区别于核心动词,因其与核心谓词的词性标注形式近似,皆以“VP” 为起始标识,如,动名词“doing”被标识为“(VP (VBG doing)”,动词不定式被标识为“(VP (TO to)”。通过预处理,减少噪音,提升准确度;

(3)以“SBAR”(从句标识符号)、“,”及“CC”(并

列连接标识符号）等符号为特征标识，进行长句的预分割处理；

（4）基于 SAO 结构，对预分割结果进行验证与合并，保证每个分段中只有一个核心谓词，并输出最终的分割结果。

如句子“Then, two models of damping in a tall building, the artificial neural network (ANN) model and the auto-regressive (AR) model, are established by employing ANN and AR methods, and used to predict the damping values at high amplitude level, which are difficult to obtain form field measurements.”（来自论文题名为“Damping in buildings: its neural network model and AR model”的摘要）基于本文上述方法，初始分段为 8 段，验证合并后，结果为 3 段，中间一段为“and used to predict the damping values at high amplitude level”。

2.3　依存句法特征分析的实体及其关系抽取

根据 Standford TypedDependency 依赖关系函数，对基于 SAO 结构的语义片段进行计算，找出核心谓词及与其语义相关的主语和宾语。以上述例句的前两个片段为例，对科研命名实体关系抽取流程进行说明，主要包括：实体抽取、实体关系识别、依存句法特征分析等过程，其所对应的依赖分析、块分析、词性标注示例如图 2 所示：

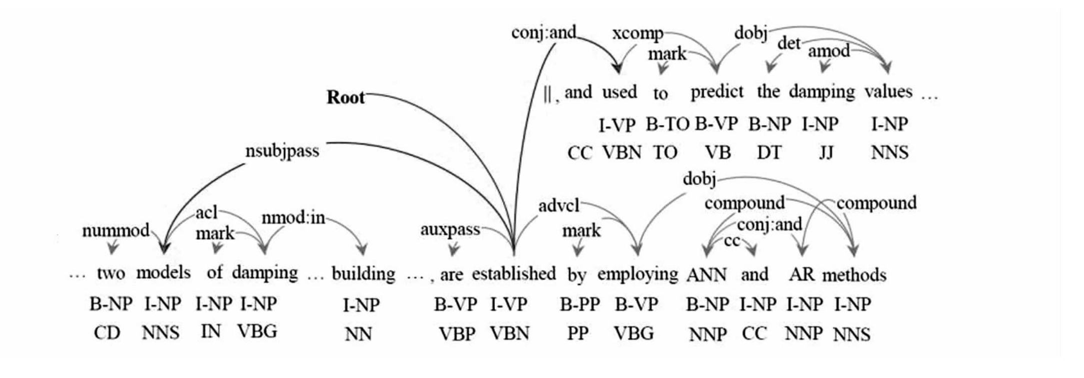


图 2　基于 Standford nlp 工具的句子依赖分析、块分析、词性标注结果示例

2.3.1　实体抽取

依据词性标注结果和基础句型，围绕谓词整理实体抽取规则，并实现实体抽取。实现过程主要包含如下两个步骤：①利用模式匹配工具（Tregex）[23]，通过执行模式“NP ! << NP”对最小 NP 块进行识别。在语法树中最小的 NP 块（noun phrase，名词词组），被认为是语义处理的最小单元。处理结果：“two models of damping in a tall building”“ANN and AR methods”“the damping values”等 NP 块被分别当作一个独立的块。②利用规则“A established by B”，匹配抽取实体对象 A 和 B，第一个语义片段的候选三元组结果为：（two models of damping in a tall building; be established by; ANN and AR methods），实体类型分别为“question”和“method”；利用规则“A used to B”，匹配抽取实体对象 A 和 B，第二语义片段的候选三元组结果为：（two models of damping in a tall building; be used to; the damping values），而实体类型分别为“method”和“question”。

2.3.2　实体关系识别

实体关系的抽取，主要是对实体对象之间的整体部分、组成、施事、因果关系进行的识别。本文借鉴蒋婷[24]对学术文献中常见实体关系类型的归纳，利用 WordNet[25]工具对主要关系类型中的具有术语类别依赖的动词（谓词）进行补充和扩展。结合科技文献特点，在 SVOA 模型[22]中增加辅助词（“to”“for”“with”“as”“in”等）的依存规则和抽取规则，提升科研命名实体的识别能力，如：use Method for Question 等。

2.3.3　基于句法依存分析的精准识别

值得注意的是，第二个语义片段的实体对象抽取过程中，其本身并不存在实体对象 A，本文利用句法依存关系分析与关系链计算进行关联识别。从图 2 中可以看出，“used-> established-> models”的依赖关系链。依赖关系识别模式为：“({} = object > conj:and {lemma:used} = {}) > nsubjpass {} = subject”，通过模式匹配找到“used”的关联实体对象 A 为“two models of damping in a tall building”。同理，基于依赖关系链分析还可以解决实体对象的共指，识别语义相关实体，实现实体聚类合并等。

另一个值得注意的是经过最小 NP 块合并及模式匹配识别后，得到的实体候选三元组中实体要素可能包含多个实体，还需要依据临近原则或句法依存分析结果查找其对应的合理的实体关系。例如：“Feed-For-

ward Back-Propagation Artificial Neural Network (FFBP-ANN) trained with Levenberg-Marquardt algorithm is used for estimation of different performance parameters of CMPA."句子中,针对"estimation of different performance parameters of CMPA"问题,初始抽取到的对应的方法实体为"Feed-Forward Back-Propagation Artificial Neural Network (FFBP-ANN) trained with Levenberg-Marquardt algorithm",但"Feed-Forward Back-Propagation Artificial Neural Network (FFBP-ANN) trained with Levenberg-Marquardt algorithm"包含了两个实体"Feed-Forward Back-Propagation Artificial Neural Network"和"Levenberg-Marquardt algorithm",根据依存关系链,最终的方法实体应是"Feed-Forward Back-Propagation Artificial Neural Network"而不是"Levenberg-Marquardt algorithm"。其依存关系计算的部分结果如下:[…, nsubjpass(used-14, Network-5), appos(Network-5, FFBP-ANN-7), acl(Network-5, trained-9), …, nmod:with(trained-9, algorithm-12), auxpass(used-14, is-13), root(ROOT-0, used-14)…]。即,通过依存分析结果可以看出,"root(ROOT-0, used-14)"说明句子的核心词为"used";"nsubjpass(used-14, Network-5)"说明"Network"是"used"的主语("nsubjpass"标识含义为被动的名词主语);"nmod:with(trained-9, algorithm-12)"说明"algorithm"与"trained"通过"with"组成复合名词("nmod"标识含义为复合名词修饰);"acl(Network-5, trained-9)"说明复合名词修饰"Network"。为了在后续的实体及关系识别过程中复用这个分析结果,本文将"Network"和"used"的依存关系链定义为"ner.dep_nsubjpass_identifier()"。同理,归纳总结核心谓词与辅助词之间、共指词之间的依存关系链,为常见依存关系链建模,形成依存关系链判别模型,实现接口复用以及对科研命名实体的精准识别。

3 实证研究

3.1 实验设计和实验步骤

本文从微软学术数据库中提取 *Artificial Intelligence* 期刊2016年发表的被引用量Top10的论文的文摘,作为实验数据。使用Ollie-app-latest.jar、Reverb-latest.jar、Stanford-corenlp-3.9.2.jar、Stanford-tregex-3.9.2.jar为主要开发工具,JDK1.8为开发环境。利用句法标注和依存关系链分析,构造科研命名实体抽取规则模型和依存关系模型,对科研文本中的重要术语及其关系进行识别和揭示。然后,通过将本文的识别算法同Ollie[12]、Reverb[9]算法及人工标注的结果进行对比分析,证明文本提出的算法在科研命名实体及其关系识别中的有效性。实验过程主要包括如下几个方面:

(1)利用基础的自然语言处理工具,设计科研命名实体识别算法,归纳整理常见的句法模型和依存关系模型,构建本文的科研命名实体识别原型系统;

(2)通过人工标注实验数据中的重要术语、实体及其关系,作为对比的基准数据;

(3)将Ollie、Reverb开放信息抽取(Information Extraction: IE)工具作为科研命名实体识别的对比算法,获得识别结果。

3.2 开放信息抽取(IE)工具

Reverb[9]与Ollie[12]是由美国华盛顿大学推出的开放信息抽取工具,通过识别任意句子中的实体关系,完成实体及其关系的提取。Reverb是早期作品,主要抽取基于动词的实体关系,即SAO结构中,通过A来提取S和O。Ollie是Reverb的升级版本,在关系识别模式和上下文信息辅助判别等方面进行了较大改进,进而推出的新一代信息抽取工具。

在关系识别模式方面,Ollie加入了以名词、形容词为关联介质的关系判别模式。如:"Microsoft co-founder Bill Gates spoke at..."的抽取结果为(Bill Gates; be co-founder of; Microsoft),其中"co-founder"即为以名词为关联介质的关系判别。上下文信息辅助判别方面,Ollie使用属性和子句修饰符等信息,提高抽取质量。如:"Early astronomers believed that the earth is the center of the universe."的抽取结果为((the earth; be the center of; the universe), AttributedTo believe; Early astronomers),属性信息说明结论与简单抽取的信息是相反的。如:"If he wins five key states, Romney will be elected President."的抽取结果为((Romney; will be elected; President) ClausalModifier if; he wins five key states),子句修饰符提供了更多的信息。

3.3 实验结果分析

按照实验步骤利用本文提出的算法、人工标注、Reverb、Ollie算法分别对实验数据进行处理,并分为两个部分对实验结果进行分析:一是对整体实验进行分析,二是结合单篇文摘实例进行分析。

3.3.1 整体实验结果分析

将本文提出算法的识别结果同人工标注的基准数据和Ollie/Reverb算法识别的结果进行精确匹配和近似匹配。精确匹配是指直接和基准数据进行一对一的

匹配。近似匹配是指基于语义相关度的匹配,认为与基准数据中的实体词义高度相近的术语实体也可以被看作是正确识别结果。例如,表 2 中,人工标注基准实体为"learning algorithms",如实验算法的识别结果为"Conventional online learning algorithms",则认为是正确识别的。

在识别结果评价指标的选择上,本文采用准确率和召回率作为实体及其关系识别效果的评价指标,准确率和召回率公式分别为:

$$Precision = R_a \cap R_h / R_a \qquad 公式(1)$$

$$Recall = R_a \cap R_h / R_h \qquad 公式(2)$$

其中 *Pewxiaion* 为准确率指标,*Recall* 为召回率指标,R_a 为基于算法抽取的实体集合中的实体个数,R_h 为基于人工判别核准的数据集合中的实体个数,$R_a \cap R_h$ 表示抽取结果与人工判别结果可以匹配的实体个数。对比结果如表 1 所示:

表 1 Ollie、Reverb 及本文识别算法结果对比

基准数据	指标	匹配模式	识别方法		
			Ollie	Reverb	本文算法
人工标注的实体	准确率	近似匹配	71.1%	57.8%	**76.6%**
		精确匹配	56.7%	48.2%	**66.2%**
	召回率	近似匹配	**74.2%**	51.6%	63.4%
		精确匹配	**59.1%**	43.0%	54.8%
人工标注的实体关系	准确率	近似匹配	58.7%	38.7%	**78.0%**
		精确匹配	50.8%	32.3%	**70.0%**
	召回率	近似匹配	71.2%	46.2%	**75.0%**
		精确匹配	61.5%	38.5%	**67.3%**

通过对比结果看,本文提出的算法,除了在实体识别方面的召回率较低之外,在实体识别的准确率、实体关系识别的准确率和召回率上,比 Ollie 和 Reverb 均有优势,近似匹配的实体识别准确率达到 76.6%,近似匹配的实体关系识别准确率达到 78%,召回率达到 75%。句法特征的依存关系分析与建模在命名实体识别的准确度上起到了关键作用。

文章对科研命名实体识别算法中的实体识别召回率较低的原因,进行了简单分析:①本文识别算法本质上是一种基于规则的算法,面临所有基于规则的算法需要面对的问题——算法并不可能对所有的规则进行罗列;②Ollie 工具中包含基于动词、名词、形容词为关联介质的关系判别与识别,在实体识别过程中约束限制相对较少,所以,会有"we""paper""result"等没有实际意义的辅助的实体,抽取到的相关实体对也相对较多,导致召回率较高;③本文算法增加了依存关系特征分析模块,在提升准确度的同时也在一定程度上降低了召回率。

3.3.2 单篇文摘实例分析

通过整体分析,在一定程度上证明了本文提出的算法的有效性。下面对实验数据中具体的单篇文摘实例进行分析,并与 Ollie、Reverb 工具识别效果进行对比,进一步证明本文算法的有效性。实验实例标题为"One-pass AUC optimization"的文摘数据,部分识别结果见表 2。

相较 Ollie、Reverb 算法,本文算法优势体现在:

(1)基于动词与辅助词组合模型的关系识别效果较好,如原句:"To efficiently handle high-dimensional data, we develop two deterministic algorithms that approximate the covariance matrices." Ollie/Reverb 识别算法只能识别基于动词"develop"的实体关系(we; develop; two deterministic algorithms)。本文算法还可以识别出基于"develop to"组合模型的实体关系(two deterministic algorithms; be developed to; high-dimensional data)。从句子本身含义来看,科研文献中的实体及其关系抽取的任务目标,更希望是"developed A to handle B"中的科研命名实体 A 和 B,表 2 中的粗体字。

当出现多个辅助词时,如:"Their friendship developed through their shared interest in the Arts.",本文算法也可通过依存关系链模型,判别"through"和"in"和句子中核心谓词"developed"的依存关系,保证识别精准。

(2)基于 SAO 结构的实体识别过程中,增加了噪音处理,有效提升了识别精度。Ollie、Reverb 等测试工具的错误率相对较高,表 2 中基于 Ollie 算法的结果只提取了置信度 >0.5 的结果。在原始结果中,将"Conventional online learning algorithms …"识别为 0.436:(Conventional online; be going only once through; training data),明显是不对的。动名词"learning"的标识没有预处理好,导致名词组"Conventional online learning algorithms"被分割。

(3)基于动词之外的关联介质的关系识别效果较好,如原句:"We present a multilingual Named Entity Recognition approach based on a robust and general set of features across languages and datasets." 本文算法与 Ollie 算法都可以基于"based"(过去分词做定语修饰"approach"),识别出关系(a multilingual Named Entity Recognition approach; be based on; a robust and general set of features)。本文主要增加了现在分词、过去分词

表 2 基于 Ollie、Reverb 及本文识别算法的实体识别结果

文摘编号	1
文章来源	GAO W, WANG L, JIN R, et al. One-pass AUC optimization[J]. Artificial intelligence, 2016, 236:1－29.
结果格式	关系置信度:(S; A; O)[abbr]。
基于 Ollie 工具的重要关系识别结果	0.936:(learning algorithms; **cannot be applied directly to**; one-pass AUC optimization)[**enabler** = because AUC is measured by a sum of losses defined over pairs of instances from different classes] 0.911:(AUC; is measured by; a sum of losses) 0.831:(losses; be defined over; pairs of instances) 0.818:(AUC; is; an important performance measure that has been used in diverse tasks, such as…, etc) 0.756:(We; develop; a regression-based algorithm which only needs to maintain … training data) 0.741:(we; focus on; one-pass AUC optimization) **0.739:(we; develop; two deterministic algorithms)[enabler = this work, we focus on … To efficiently handle high-dimensional data]** 0.706:(losses; be defined from; different classes)
基于 Reverb 工具的重要关系识别结果	0.3818:(an important performance measure; has been used in; diverse tasks) 0.5853:(algorithms; **cannot be applied directly to**; one-pass AUC optimization) 0.1938:(AUC; is; an important performance measure) 0.1415:(AUC; is measured by; a sum of losses) 0.0989:(one-pass AUC optimization; requires going through; training data) 0.0713:(we; focus on; one-pass AUC optimization) 0.0544:(We; develop; a regression-based algorithm) 0.0573:(a regression-based algorithm; only needs to maintain; the first and second-order statistics of training data) **0.0176:(we; develop; two deterministic algorithms)**
本文的识别算法与接口的识别结果	0.855556:(an important performance measure; be used in; diverse tasks) 0.847826:(AUC; be measured by; a sum of losses) **0.619565:(two deterministic algorithms; be developed to; high-dimensional data)** 0.484375:(AUC; be; an important performance measure) 0.48:(Conventional online learning algorithms; **not be applied to**; one-pass AUC optimization) 0.0:(We; develop; a regression-based algorithm) **0.0:(we; develop; two deterministic algorithms)**

注:置信度计算:Ollie,Reverb 使用逻辑回归算法;本文使用距离相似度

和形容词为关联介质的实体识别模型,同时增加了这些关系连接词与辅助词("to""for""with""as""in"等)的依存关系链判别模型。

3.3.3 科研命名实体识别算法错误的原因分析

经分析发现,造成本文识别算法的错误的主要原因可归纳为如下 4 个方面:

(1)由于严重依赖于词性标注和依存关系解析器,因此由词性标注和依存关系解析错误引起的识别错误,占比较大,约 46%;如:"The results display the potential of algorithm selection to achieve significant performance improvements across a broad range of problems and algorithms"句子中,将"display"标注为"(VP (NN display)"是明显的错误;又如:"The optimization objective we study asks to minimize the expected total cost of reaching a state in the target set, while ensuring that the target set is reached almost surely."句子中,将"the target set is reached"标注为"(S (NP (DT the) (NN target)) (VP (VBD set)))))) (VP (VBZ is) (VP (VBN reached)",其中"set"被标注为动词是明显错误。前一个错误可后期修正,而后一个错误修正的难度很大,需要依赖 Standford nlp 工具的升级了。

(2)缺少"异常规则限制"模板而引起的识别错误约 20%;如:"Unfortunately, it is relatively easy to develop sophisticated models to help reduce the error of estimation by a few percent"中的"to help reduce"从句意上讲也可写成"to reduce",不会有大的歧义,但从标注结果看,与常见不定式结构略有不同,需要特殊处理;又如:"During this research a prototype of a 3D cadastre was developed."的常见表达为"During this research, a prototype of a 3D cadastre was developed."增加一个","会使句子结构更清晰,也需要添加特殊的处理规则,来修正错误。

(3)上下文共指链识别错误而引起的识别错误约 12%;上下文共指链复杂,共指链判别模型没有覆盖到的情况将出现错误,如:"Our results were satisfactory and were compared with those obtained by a learning system based on Self-Organizing Maps."中的"those"被标注为限定词(DT),共指词为"results";"We also take the opportunity to clarify some properties of the semidefinite relaxation, were it to be used for an actual nonconvex

problem in this area."中的"it"被标注为人称指代词(PRP),共指词为"properties";"In the present study, a time series neuro-fuzzy model is proposed that is capable of exploiting the strengths of traditional time series approaches."中的"that"被标注为定语从句连词(SBAR IN),共指词为"model",等等。共指链判别模型类型多,规则多。该错误可以通过共指依存关系链判别模型的丰富进一步优化,减少错误。

(4)其他错误约 22%。如特殊的复杂句型、特殊的复合词等等。

4 结语

实体及其关系抽取在许多自然语言处理任务中被证明是有用的[21]。本文针对长句的噪音问题和科研命名实体抽取的特殊性,结合词性标注和句法依存分析,对利用模式匹配提取实体关系三元组的实体抽取模型进行改进,并通过实例对以上改进方案进行了验证。新的改进措施对进一步研究科研命名实体及其关系的精准识别和抽取,具有一定的理论和实践参考价值。文章主要创新点与贡献包括:①通过对长句处理的改进,明确目标文本的语义结构,助力实体识别精准性的提升;②通过核心谓词及其相关辅助词的依存分析,对依存关系链建模,有力提升对科研命名实体的识别和抽取效果;③以科研问题及其相关解决方法为例,给出了科研命名实体识别的基本思路,有助于科研问答应用的实现。

同时也应看到,本文试验数据集合较少,缺少大数据集上的应用测试,同时核心谓词的分类整理与扩展完善、基于谓词的实体抽取模板的积累完善、以名词/形容词等为关联词的科研命名实体识别模型的补充完善、置信度计算等,皆是下一步继续努力的方向。

参考文献:

[1] 徐芬, 王挺, 陈火旺. 基于 SVM 方法的中文实体关系抽取[C]//大连理工大学,清华大学智能技术与系统国家重点实验室. 内容计算的研究与应用前沿——第九届全国计算语言学学术会议论文集. 大连理工大学,清华大学智能技术与系统国家重点实验室:中国中文信息学会, 2007:497-502.

[2] KAMBHATLA N. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations[C]//Proceedings of the ACL 2004 on interactive poster and demonstration sessions. Stroudsburg: ACL, 2004:1-4.

[3] 郭喜跃, 何婷婷, 胡小华, 等. 基于句法语义特征的中文实体关系抽取[J]. 中文信息学报, 2014, 28(6):183-189.

[4] 甘丽新, 万常选, 刘德喜, 等. 基于句法语义特征的中文实体关系抽取[J]. 计算机研究与发展, 2016, 53(2):284-302.

[5] LI H, WU X, LI Z, et al. A relation extraction method of Chinese named entities based on location and semantic features[J]. Applied intelligence, 2013, 38(1):1-15.

[6] 奚斌, 钱龙华, 周国栋, 等. 语言学组合特征在语义关系抽取中的应用[J]. 中文信息学报, 2008, 22(3):44-50.

[7] BANKO M, CAFARELLA M J, SODERLAND S, et al. Open information extraction from the Web[C]//Proceedings of the 20th international joint conference on artificial intelligence, Hyderabad, India. San Francisco: Morgan Kaufmann Publishers Inc., 2007: 2670-2676.

[8] FADER A, SODERLAND S, ETZIONI O. Identifying Relations for Open Information Extraction[C]//Proceedings of the 2011 conference on empirical methods in natural language processing. Stroudsburg: ACL, 2011:1535-1545.

[9] ETZIONI O, FADER A, CHRISTENSEN J, et al. Open information extraction: the second generation[C]//Proceedings of conference on artificial intelligence. Palo Alto: AAAI Press, 2011:3-10.

[10] WU F, WELD D S. Open Information Extraction Using Wikipedia[C]//Proceedings of the 48th annual meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2010: 118-127.

[11] AKBIK A, LÖSER A. KrakeN: N-ary facts in open information extraction[C]//Proceedings of the joint workshop on automatic knowledge base construction and web-scale knowledge extraction. Stroudsburg: ACL, 2012:52-56.

[12] SCHMITZ M, BART R, SODERL S, et al. Open language learning for information extraction[C]//Proceedings of the conference on empirical methods in natural language processing and computational natural language learning. Stroudsburg: ACL, 2012:523-534.

[13] MAUSAM M. Open information extraction systems and downstream applications[C]//Proceedings of the twenty-fifth international joint conference on artificial intelligence. Palo Alto: AAAI Press, 2016: 4074-4077.

[14] 武文雅, 陈钰枫, 徐金安, 等. 中文实体关系抽取研究综述[J]. 计算机与现代化, 2018(8):21-27.

[15] 唐敏. 基于深度学习的中文实体关系抽取方法研究[D]. 成都:西南交通大学, 2018.

[16] LIN Y, LIU Z, SUN M. Neural relation extraction with multi-lingual attention[C]//Proceedings of the 55th annual meeting of the Association for Computational Linguistics. Vancouver:ACL, 2017: 34-43.

[17] ILEVBARE I M, PROBERT D, PHAAL R. A review of TRIZ, and its benefits and challenges in practice[J]. Technovation,

2013, 33(2):30－37.

[18] CHOI S, YOON J, KIM K, et al. SAO network analysis of patents for technology trends identification: a case study of polymer electrolyte membrane technology in proton exchange membrane fuel cells [J]. Scientometrics, 2011, 88(3):863－883.

[19] 郭俊芳，汪雪锋，邱鹏君，等. 基于SAO分析的技术路线图构建研究[J]. 科学学研究，2014(7):976－981.

[20] 汪雪锋，邱鹏君，付芸. 一种新型技术路线图构建研究——基于SAO结构信息[J]. 科学学研究，2015(8):1134－1140.

[21] ANGELI G, PREMKUMAR M J, MANNING C D. leveraging linguistic structure for open domain information extraction[C]//Proceedings of the 53rd annual meeting of the Association for Computational Linguistics and the 7th international joint conference on natural language processing. Stroudsburg: ACL, 2015: 344－354.

[22] CORRO L D, GEMULLA R. ClausIE: Clause-based open information extraction[C]//Proceedings of the 22nd international conference on World Wide Web. New York: ACM, 2013:355－366.

[23] Tregex, Tsurgeon and Semgrex[EB/OL]. [2019－09－17]. https://nlp.stanford.edu/software/tregex.shtml.

[24] 蒋婷，孙建军. 学术资源本体非等级关系抽取研究[J]. 图书情报工作，2016, 60(20):112－122.

[25] What is WordNet? [EB/OL]. [2019－09－17]. https://wordnet.princeton.edu/.

作者贡献说明：
赵华茗：选题制定、方法设计、算法测试、论文撰写；
钱力：研究框架设计，论文修改；
余丽：收集、整理语料库数据，论文修改；

A Research Entity Recognition Algorithm Based on Dependency Parsing

Zhao Huaming[1] Qian Li[1,2] Yu Li[1]

[1] National Science Library, Chinese Academy of Sciences, Beijing 100190

[2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

Abstract: [**Purpose/significance**] To explore the recognition and extraction of research entities and their relationships, improve their recognition effect in complex situations such as long sentences, and provide reference for further application. [**Method/process**] Based on the analysis of dependency syntactic features, a method for recognizing and extracting research entity relations was proposed, which includes: POS tagging of the target text using Standford Tagger tool; based on annotation results, the target text was divided into semantic segments of structure specification around the core predicate and SAO structure; through dependency parsing, we can find out the subject and object related to the core predicate and form a triple of entities, relationships and entities. [**Result/conclusion**] This method is compared with Ollie and Reverb mainstream algorithm. Experiments show that this method can effectively improve the accuracy of scientific entity recognition.

Keywords: dependency parsing research entity entity recognition relation extraction